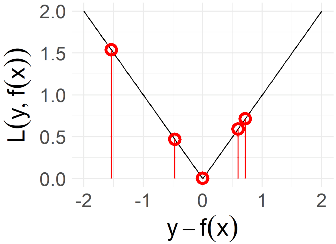
Introduction to Machine Learning

Regression Losses: L1-loss



Learning goals

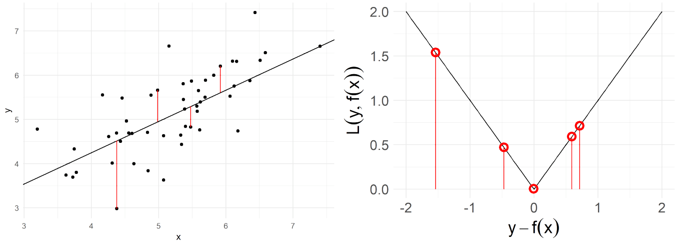
- Derive the risk minimizer of the L1-loss
- Derive the optimal constant model for the L1-loss

L1-LOSS

The L1 loss is defined as

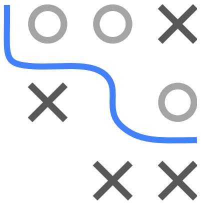
$$L\left(y, f(\mathbf{x})\right) = |y - f(\mathbf{x})|$$

- More robust than $L2$, outliers in y are less problematic.
- Analytical properties: convex, not differentiable for $y = f(\boldsymbol{x})$ (optimization becomes harder).



L1-LOSS: RISK MINIMIZER

We calculate the (true) risk for the $L1$-Loss $L\left(y, f(\mathbf{x})\right) = |y - f(\mathbf{x})|$ with unrestricted $\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y}\}$.



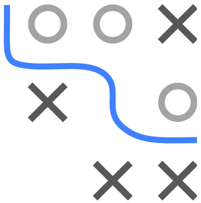
- We use the law of total expectation

$$\mathcal{R}(f) = \mathbb{E}_x \left[\mathbb{E}_{y|x}\left[|y - f(\mathbf{x})| \;|\mathbf{x} = \mathbf{x}\right]\right].$$

- As the functional form of f is not restricted, we can just optimize point-wise at any point $\mathbf{x} = \mathbf{x}$. The best prediction at $\mathbf{x} = \mathbf{x}$ is then

$$\hat{f}(\mathbf{x}) = \text{argmin}_c \mathbb{E}_{y|x}\left[|y - c|\right] = \text{med}_{y|x}\left[y \mid \mathbf{x}\right].$$

L1-LOSS: RISK MINIMIZER / 2



Proof: Let $p(y)$ be the density function of y. Then:

$$
\begin{aligned}
\text{argmin}_c \mathbb{E}\left[|y - c|\right] &= \text{argmin}_c \int_{-\infty}^{\infty} |y - c| \, p(y) \text{d}y \\
&= \text{argmin}_c \int_{-\infty}^{c} -(y - c) \, p(y) \, \text{d}y + \int_{c}^{\infty} (y - c) \, p(y) \, \text{d}y
\end{aligned}
$$

We now compute the derivative of the above term and set it to 0

$$
\begin{aligned}
0 &= \frac{\partial}{\partial c} \left(\int_{-\infty}^{c} -(y - c) \, p(y) \, \text{d}y + \int_{c}^{\infty} (y - c) \, p(y) \, \text{d}y \right) \\
&\overset{*\text{Leibniz}}{=} \int_{-\infty}^{c} p(y) \, \text{d}y - \int_{c}^{\infty} p(y) \, \text{d}y = \mathbb{P}_y(y \leq c) - (1 - \mathbb{P}_y(y \leq c)) \\
&= 2 \cdot \mathbb{P}_y(y \leq c) - 1 \\
\Leftrightarrow 0.5 &= \mathbb{P}_y(y \leq c),
\end{aligned}
$$

which yields $c = \text{med}_y(y)$.

L1-LOSS: RISK MINIMIZER / 3

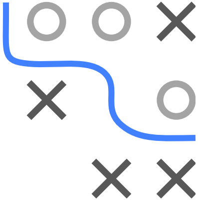
* **Note** that since we are computing the derivative w.r.t. the integration boundaries, we need to use Leibniz integration rule

$$\begin{aligned}
\frac{\partial}{\partial c}\left(\int_a^c g(c,y)\,\mathrm{d}y\right) &= g(c,c) + \int_a^c \frac{\partial}{\partial c} g(c,y)\,\mathrm{d}y \\
\frac{\partial}{\partial c}\left(\int_c^a g(c,y)\,\mathrm{d}y\right) &= -g(c,c) + \int_c^a \frac{\partial}{\partial c} g(c,y)\,\mathrm{d}y
\end{aligned}$$

We get

$$\begin{aligned}
& \frac{\partial}{\partial c}\left(\int_{-\infty}^c -(y-c)\,p(y)\,\mathrm{d}y + \int_c^\infty (y-c)\,p(y)\,\mathrm{d}y\right) \\
= \; & \frac{\partial}{\partial c}\left(\int_{-\infty}^c \underbrace{-(y-c)\,p(y)}_{g_1(c,y)}\,\mathrm{d}y\right) + \frac{\partial}{\partial c}\left(\int_c^\infty \underbrace{(y-c)\,p(y)}_{g_2(c,y)}\,\mathrm{d}y\right) \\
= \; & \underbrace{g_1(c,c)}_{=0} + \int_{-\infty}^c \frac{\partial}{\partial c}\left(-(y-c)\right)\,p(y)\,\mathrm{d}y - \underbrace{g_2(c,c)}_{=0} + \int_c^\infty \frac{\partial}{\partial c}(y-c)\,p(y)\,\mathrm{d}y \\
= \; & \int_{-\infty}^c p(y)\,\mathrm{d}y + \int_c^\infty -\,p(y)\,\mathrm{d}y.
\end{aligned}$$

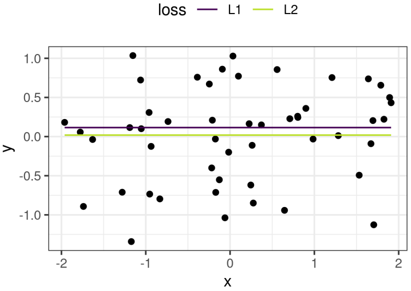
L1-LOSS: OPTIMAL CONSTANT MODEL



The optimal constant model in terms of the theoretical risk for the L1 loss is the median over y:

$$f(\mathbf{x}) \quad = \quad \text{med}_{y|x}\left[y \mid \mathbf{x}\right] \overset{\text{drop } \mathbf{x}}{=} \text{med}_y\left[y\right]$$

The optimizer of the empirical risk is $\text{med}(y^{(i)})$ over $y^{(i)}$, which is the empirical estimate for $\text{med}_y\left[y\right]$.



L1-LOSS: OPTIMAL CONSTANT MODEL / 2

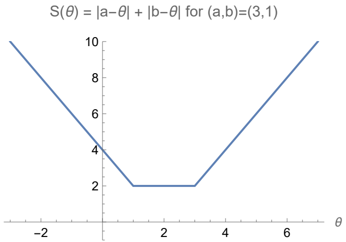
Proof:

- Firstly note that for $n = 1$ the median $\hat{\theta} = \text{med}(y^{(i)}) = y^{(1)}$ obviously minimizes the emp. risk $\mathcal{R}_{\text{emp}}$ using the $L1$ loss.
- Hence let $n > 1$ in the following For $a, b \in \mathbb{R}$, define

$$S_{a,b} : \mathbb{R} \to \mathbb{R}_0^+, \theta \mapsto |a - \theta| + |b - \theta|$$

Any $\hat{\theta} \in [a, b]$ minimizes $S_{a,b}(\theta)$, because it holds that

$$S_{a,b}(\theta) = \begin{cases} |a - b|, & \text{for } \theta \in [a, b] \\ |a - b| + 2 \cdot \min\{|a - \theta|, |b - \theta|\}, & \text{otherwise.} \end{cases}$$



L1-LOSS: OPTIMAL CONSTANT MODEL / 3

W.l.o.g. assume now that all $y^{(i)}$ are sorted in increasing order.
Let us define $i_{\text{max}} = n/2$ for n even and $i_{\text{max}} = (n-1)/2$ for n odd and consider the intervals

$$\mathcal{I}_i := [y^{(i)}, y^{(n+1-i)}], i \in \{1, ..., i_{\text{max}}\}.$$

By construction $\mathcal{I}_{j+1} \subseteq \mathcal{I}_j$ for $j \in \{1, \ldots, i_{\text{max}} - 1\}$ and $\mathcal{I}_{i_{\text{max}}} \subseteq \mathcal{I}_i$.
With this, $\mathcal{R}_{\text{emp}}$ can be expressed as

$$\begin{aligned}
\mathcal{R}_{\text{emp}}(\theta) &= \sum_{i=1}^{n} L(y^{(i)}, \theta) = \sum_{i=1}^{n} |y^{(i)} - \theta| \\
&= \underbrace{|y^{(1)} - \theta| + |y^{(n)} - \theta|}_{=S_{y^{(1)}, y^{(n)}}(\theta)} + \underbrace{|y^{(2)} - \theta| + |y^{(n-1)} - \theta|}_{=S_{y^{(2)}, y^{(n-1)}}(\theta)} + \ldots \\
&= \begin{cases} \sum_{i=1}^{i_{\text{max}}} S_{y^{(i)}, y^{(n+1-i)}}(\theta) & \text{for } n \text{ is even} \\ \sum_{i=1}^{i_{\text{max}}} \left(S_{y^{(i)}, y^{(n+1-i)}}(\theta)\right) + |y^{((n+1)/2)} - \theta| & \text{for } n \text{ is odd.} \end{cases}
\end{aligned}$$

L1-LOSS: OPTIMAL CONSTANT MODEL / 4

From this follows that

- for "n is even": $\hat{\theta} \in \mathcal{I}_{i_{\max}} = [y^{(n/2)}, y^{(n/2+1)}]$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\}$ $\Rightarrow$ it minimizes $\mathcal{R}_{\text{emp}}$,
- for "n is odd": $\hat{\theta} = y^{(n+1)/2} \in \mathcal{I}_{i_{\max}}$ minimizes S_i for all $i \in \{1, \dots, i_{\max}\}$ and it's minimal for $|y^{((n+1)/2)} - \theta|$ $\Rightarrow$ it minimizes $\mathcal{R}_{\text{emp}}$.

Since the median fulfills these conditions, we can conclude that it minimizes the $L1$ loss.